HW2_STAT5000

Xingyu Chen

September 10, 2021

Due on Friday September 10, 2021 @ 11:59 PM through Canvas. Covers exploratory data analysis and intro to probability. Instructions for "theoretical" questions: Answer all of the following questions. The theoretical problems should be neatly numbered, written out or typed, and self-graded. This homework has 20 questions or lettered question parts, totaling 40 points.

Total Score: 38/40

1 Theoretical Questions

1.

(a) Let a and b be constants and let $y_i = ax_i + b$ for $i = 1, ..., n$. What are the relationships between $\overline{x}$ and $\overline{y}$ and between s_x^2 and s_y^2? Q1: 2/2

Answer:

We have $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
meanwhile $\overline{y} = \frac{1}{n}(y_1 + y_2 + ... + y_n)$
$= \frac{1}{n}(ax_1 + b + ax_2 + b + ... + ax_i + b)$
$= \frac{1}{n}[a(x_1 + x_2 + ... + x_n) + n * b]$
$= a * \frac{1}{n}\sum_{i=1}^{n} x_i + b$
$= a\overline{x} + b.$
We have $S_x^2 = \frac{1}{n-1} * \sum_{i=1}^{n}(x_i - \overline{x})^2$
meanwhile $S_y^2 = \frac{1}{n-1} * \sum_{i=1}^{n}(y_i - \overline{y})^2$
$= \frac{1}{n-1} * \sum_{i=1}^{n}(ax_i + b - (a\overline{x} + b))^2$
$= \frac{1}{n-1} * \sum_{i=1}^{n} a^2(x_i - \overline{x})^2$
$= \frac{a^2}{n-1} * \sum_{i=1}^{n}(x_i - \overline{x})^2$
$= a^2 S_x^2.$

(b) A sample of temperatures for initiating a certain chemical reaction yielded a sample average of 87.3 degrees Celsius and a sample standard deviation of 1.04. What are the sample average and standard deviation measured in Fahrenheit [HINT: F = 9/5C + 32]? Q2: 2/2

Answer:

x = ^oC, y = ^oF.

$\overline{y} = 9/5\ (87.3) + 32 = 189.14,$
$S_y = \sqrt{S_y^2} = \sqrt{(9/5)^2(1.04)^2} = \sqrt{3.5044} = 1.872.$

2. Let $\overline{x}_n$ and s_n^2 denote the sample mean and variance for the sample $x_1, ..., x_n$ and let $\overline{x}_{n+1}$ and s_{n+1}^2 denote these quantities when an additional observation x_{n+1} is added to the sample.

(a) Show that $\overline{x}_{n+1} = \overline{x}_n + \frac{x_{n+1} - \overline{x}_n}{n+1}$ Q3: 2/2

Answer:

$\overline{x}_{n+1} = \frac{1}{n+1}\sum_{i=1}^{n+1} x_i$
$= \frac{1}{n+1}(\sum_{i=1}^{n} x_i + x_{n+1})$
$= \frac{1}{n+1}(n\overline{x}_n + x_{n+1})$
$= \frac{n\overline{x}_n + x_{n+1}}{n+1}$
$= \overline{x}_n + \frac{x_{n+1} - \overline{x}_n}{n+1}$

(b) Show that $s_{n+1}^2 = \frac{(n-1)}{n}s_n^2 + \frac{1}{n+1}(x_{n+1} - \overline{x}_n)^2$. Q4: 2/2

Answer:

We have $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$,
$s_{n+1}^2 = \frac{1}{n}\sum_{i=1}^{n+1}(x_i - \overline{x}_{n+1})^2$
$= \frac{1}{n}\sum_{i=1}^{n+1}(x_i - \overline{x}_n + \overline{x}_n - \overline{x}_{n+1})^2$
$= \frac{1}{n}\sum_{i=1}^{n+1}[(x_i - \overline{x}_n)^2 + 2(x_i - \overline{x}_n)(\overline{x}_n - \overline{x}_{n+1}) + (\overline{x}_n - \overline{x}_{n+1})^2]$
$= \frac{1}{n}[\sum_{i=1}^{n+1}(x_i - \overline{x}_n)^2 + 2(\overline{x}_n - \overline{x}_{n+1})\sum_{i=1}^{n+1}(x_i - \overline{x}_n) + (n+1)(\overline{x}_n - \overline{x}_{n+1})^2]$
$= \frac{1}{n}[\sum_{i=1}^{n+1}(x_i - \overline{x}_n)^2 + (x_{n+1} - \overline{x}_n)^2 + 2(\overline{x}_n - \overline{x}_{n+1})[(n+1)\overline{x}_{n+1} - (n+1)\overline{x}_n] + (n+1)(\overline{x}_n - \overline{x}_{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (x_{n+1} - \overline{x}_n)^2 - 2(n+1)(\overline{x}_n - \overline{x}_{n+1})^2 + (n+1)(\overline{x}_n - \overline{x}_{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (x_{n+1} - \overline{x}_n)^2 - (n+1)(\overline{x}_n - \overline{x}_{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (x_{n+1} - \overline{x}_n)^2 - (n+1)(\overline{x}_n - \frac{n\overline{x}_n + x_{n+1}}{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (x_{n+1} - \overline{x}_n)^2 - (n+1)(\frac{\overline{x}_n + x_{n+1}}{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (x_{n+1} - \overline{x}_n)^2 - \frac{1}{n+1}(\overline{x}_n - x_{n+1})^2]$
$= \frac{1}{n}[(n-1)S_n^2 + (1 - \frac{1}{n+1})(x_{n+1} - \overline{x}_n)^2]$
$= \frac{1}{n}[(n-1)S_n^2 + \frac{n}{n+1}(x_{n+1} - \overline{x}_n)^2]$
$= \frac{(n-1)}{n}s_n^2 + \frac{1}{n+1}(x_{n+1} - \overline{x}_n)^2$

(c) Why might these results be useful? Q5: 1/2

Answer:

We can see how sample mean and sample variance are changed when sample n is increase. We can also see the sample mean and sample variance are independent each other.

(d) In both (a) and (b), describe what happens as $n \to \infty$. Q6: 2/2

Answer:

The central limit theorem states that:
Given a population with a finite mean x_n and a finite non-zero variance s^2, the sampling distribution of the mean approaches a normal distribution with a mean of x_n and a

variance of s^2/n as n, the sample size, increases.

3. Three fair dice are thrown. What is the probability that a sum of 8 appears on the faces? What is the probability that a sum of 10 appears? Q7: 2/2
Answer:
$P(Sumof8) = 7C2/6^3 = 21/216 = 9.7\%$
$P(Sumof10) = [9C2 - P(1,1,8) - P(1,2,7)]/6^3 = (36-3-6)/216 = 27/216 = 12.5\%$

4. Consider randomly selecting a student at CU Boulder. Let A denote the event that the selected student has a Venmo account and let B be the event that the selected student has a Paypal account. Suppose that P(A) = 0.6 and P(B) = 0.5.

(a) Is it possible that $P(A \cap B) = 0.55$? Cite a theorem of probability in your answer. Q8: 2/2
Answer:
No.
$(A \cap B) \subseteq A$ and $(A \cap B) \subseteq B$,
Thus, $P(A \cap B) \leq min(P(A), P(B))$.
If $P(A \cap B) = 0.55$, which is greater than P(B) = 0.5, impossible.
One of Mlodinow's three laws of probability: "The probability that two events will both occur can never be greater than the probability that each will occur individually."

(b) For the remaining questions, let $P(A \cap B) = 0.3$. Compute the probability that the selected individual has at least one of the two types of accounts. Q9: 2/2
Answer:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
= 0.5+0.6-0.3
= 0.8

(c) What is the probability that the selected individual has neither type of account? Q10: 2/2
Answer:
$P'(A \cup B) = 1 - P(A \cup B)$
= 1 - 0.8
= 0.2

(d) Describe in terms of A and B the event that the selected student has Venmo but not Paypal, and then calculate the probability of this event. Q11: 2/2
Answer:
$P(AnotB) = P(A) - P(A \cap B)$
= 0.6 - 0.3
= 0.3

5. Prove that if one event A is contained in another event B (i.e., A is a subset of B) then $P(A) \leq P(B)$. [HINT: you might consider the set $B \cap A^c$ in your computation, where A^c is the complement of A.] Q12: 2/2
Answer:

$B = (B \cap A) \cup (B \cap A^c) = A \cup (B \cap A^c)$
Thus, $P(B) = P(A \cup (B \cap A^c))$
$= P(A) + P(B \cap A^c) - P(A \cap (B \cap A^c))$
$= P(A) + P(B \cap A^c) - P(\emptyset)$
$= P(A) + P(B \cap A^c)$
Since $P(B \cap A^c) \geq 0$, $P(B) \geq P(A)$.

2 Computational Questions

Instructions for "computational" questions: Your work should be neatly done and in- clude all graphs, code, and comments, labeled and in order based on the problem you are addressing. Do not put graphs in at the end, stick code in random locations, or do anything else that will make this homework difficult to read and grade.

1. Some claim that the final hours aboard the Titanic were marked by class warfare; other claim it was characterized by male chivalry. The data frame TITANIC3 from the PASWR2 package contains information pertaining to class status pclass, survival of passengers survived, and gender sex, among others. Based on the information in the dataframe:

```
library(tidyverse)
library(PASWR2)
df <- PASWR2::TITANIC3
```

(a) Determine the fraction of survivors from each passenger class. Q13: 2/2
Answer:

```
df_a <- df
df_a %>%
  group_by(pclass) %>%
  summarise(frac_survived = mean(survived) * 100)
```

pclass	frac_survived
1st	61.91950
2nd	42.96029
3rd	25.52891

(b) Compute the fraction of survivors according to class and gender. Did men in the first class or women in the third class have a higher survival rate? Q14: 2/2
Answer:

```
df_b <- df
df_b %>%
  group_by(pclass,sex) %>%
  summarise(frac_survived = mean(survived) * 100)
```

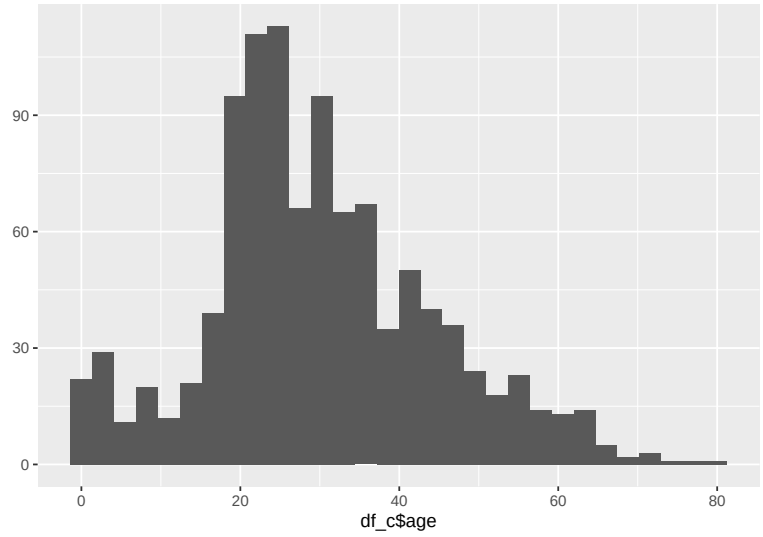
pclass	sex	frac_survived
1st	female	96.52778
1st	male	34.07821
2nd	female	88.67925
2nd	male	14.61988
3rd	female	49.07407
3rd	male	15.21298

The women in the third class have a higher survival rate.

(c) How would you characterize the distribution of age (e.g., is it symmetric, positively/negatively skewed, unimodal, multimodal)? Q15: 2/2

Answer:

```
df_c <- df
qplot(df_c$age)
```



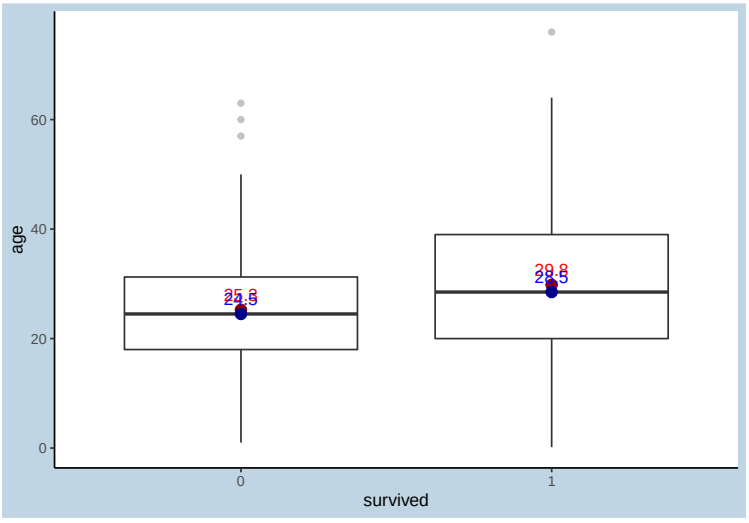
It is positively skewed.

(d) Were the median and mean ages for females who survived higher or lower than for females who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic. Q16: 2/2

Answer:

```
df_d <- df
df_d <- df_d %>%
  filter(sex == 'female')
df_d$survived <- factor(df_d$survived)
ggplot(data = df_d, aes(x = survived, y=age)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun=mean, colour="darkred", geom="point",
```

```
              hape=18, size=3,show_guide = FALSE)+
  stat_summary(fun=mean, colour="red", geom="text",
               show_guide = FALSE, vjust=-0.7,
               aes( label=round(..y.., digits=1))) +
  stat_summary(fun=median, colour="darkblue",
               geom="point", hape=18,
               size=3,show_guide = FALSE)+
  stat_summary(fun=median, colour="blue", geom="text",
               show_guide = FALSE, vjust=-0.7,
               aes( label=round(..y.., digits=1))) +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```




```
df_d %>%
  group_by(survived) %>%
  drop_na(age) %>%
  summarize(mean = mean(age), median = median(age))
```

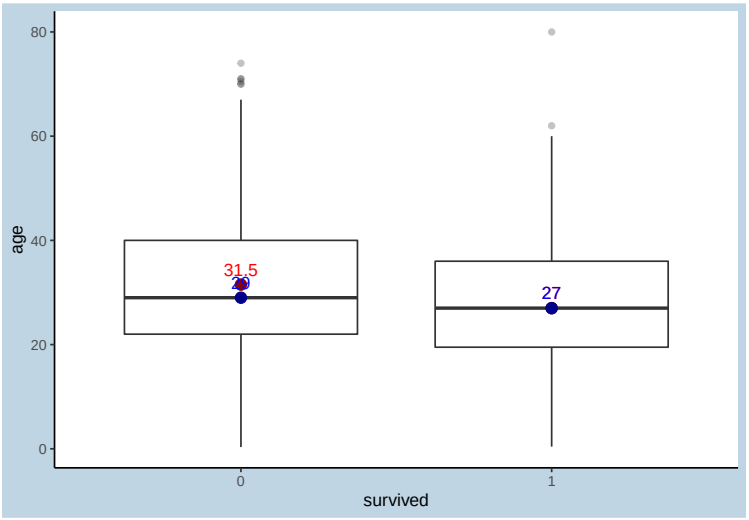
survived	mean	median
0	25.25521	24.5
1	29.81535	28.5

The median and mean ages for females who survived were higher than for females who did not survive.

(e) Were the median and mean ages for males who survived higher or lower than for males who did not survive? Report the median and mean ages as well as an appropriate measure of spread for each statistic. Q17: 2/2

Answer:

```
df_e <- df
df_e <- df_e %>%
  filter(sex == 'male')
df_e$survived <- factor(df_e$survived)
ggplot(data = df_e, aes(x = survived, y=age)) +
  geom_boxplot(alpha=0.3) +
  stat_summary(fun=mean, colour="darkred", geom="point",
               hape=18, size=3,show_guide = FALSE)+
  stat_summary(fun=mean, colour="red", geom="text",
               show.legend = FALSE, vjust=-0.7,
               aes( label=round(..y.., digits=1))) +
  stat_summary(fun=median, colour="darkblue", geom="point",
               hape=18, size=3,show.legend = FALSE)+
  stat_summary(fun=median, colour="blue", geom="text",
               show_guide = FALSE, vjust=-0.7,
               aes( label=round(..y.., digits=1))) +
  theme_classic() +
  theme(plot.background = element_rect(fill = "#BFD5E3"))
```




```
df_e %>%
  group_by(survived) %>%
  drop_na(age) %>%
  summarize(mean = mean(age), median = median(age))
```

survived	mean	median
0	31.51641	29
1	26.97778	27

The median and mean ages for males who survived were lower than for females who did not

survive.

(f) What was the age of the youngest female in the first class who survived? Q18: 2/2
Answer:

```
df_f <- df
df_f <- df_f %>%
  filter(sex == 'female' & pclass == '1st' & survived == 1) %>%
  arrange(age)
head(df_f, 1)
```

pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked	boat	body	home.dest
1st	1	Carter, Miss. Lucile Polk	female	14	1	2	113760	120	B96 B98	Southampton	4	NA	Bryn Mawr, PA

The age is 14.

(g) Do the data suggest that the final hours aboard the Titanic were characterized by class warfare, male chivalry, some combination of both, or neither? Justify your answer based on computations above, or based on other explorations of the data. Q19: 1/2
Answer:
Both class warfare and male chivalry. From computation (b), we know that female generally survival more than male and better class have higher survival rate.

2. Conduct a simulation in R to numerically illustrate the results from theoretical question 2 (a) and (b). Q20: 2/2
Answer:

```
n = sample(1:100, 20)
adding = sample(1:100, 1)
print(n)
```

```
##  [1]  5 31 90 41 63 69 61 58 67 32 52 29 94 35 48 38 74  7 40 25
```

```
print(paste0("The adding value: ", adding))
```

```
## [1] "The adding value: 42"
```

```
print(paste0("mean of n: ", mean(n)))
```

```
## [1] "mean of n: 47.95"
```

```
print(paste0("variance of n: ", var(n)))
```

```
## [1] "variance of n: 590.260526315789"
```

```
print(paste0("The mean of n with adding:", mean(c(n,adding))))
```

```
## [1] "The mean of n with adding:47.6666666666667"
```

```
print(paste0("The variance of n with adding:",var(c(n,adding))))
```

```
## [1] "The variance of n with adding:562.433333333333"
```

```
print(paste0("The formula from (a) calculate mean of n with adding:", mean(n) + (adding
```

```
## [1] "The formula from (a) calculate mean of n with adding:47.6666666666667"
```

```
print(paste0("The formula from (b) calculate variance of n with adding:", (((length(n) -
```

```
## [1] "The formula from (b) calculate variance of n with adding:562.433333333333"
```